Journal of Theoretical Biology 401 (2016) 38–42

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Inference of fitness values and putative appearance time points for evolvable self-replicating molecules from time series of occurrence frequencies in an evolution reactor

Takuyo Aita [a], Norikazu Ichihashi [a,b], Tetsuya Yomo [a,b,c,*]

[a] Exploratory Research for Advanced Technology, Japan Science and Technology Agency Yamadaoka 1-5, Suita, Osaka, Japan
[b] Department of Bioinformatic Engineering, Graduate School of Information Science and Technology Osaka University, Yamadaoka 1-5, Suita, Osaka, Japan
[c] Graduate School of Frontier Biosciences Osaka University, Yamadaoka 1-5, Suita, Osaka, Japan

HIGHLIGHTS

- We present the theoretical basis of the kinetic model used in our previous study.
- The details of inference of the putative appearance time points are presented.
- This methodology may be applicable to viruses and bacteria.

ARTICLE INFO

Article history:
Received 17 September 2015
Received in revised form
24 January 2016
Accepted 14 April 2016
Available online 16 April 2016

Keywords:
In vitro evolution
Evolvability
Evolutionary pathways
Fitness
Phylogenetic method
Quasispecies

ABSTRACT

We have established a translation-coupled RNA replication system within a cell-like compartment, and conducted an experimental evolution of the RNA molecules in the system. Then, we obtained a time series of occurrence frequencies of 91 individual genotypes through random sampling and next-generation sequencing. The time series showed a complex clonal interference and a polymorphic population called the "quasispecies". By fitting a deterministic kinetic model of evolvable simple self-replicators to the time series, we estimated the fitness value and "putative appearance time point" for each of the 91 major genotypes identified, where the putative appearance time point is defined as a certain time point at which a certain mutant genotype is supposed to appear in the deterministic kinetic model. As a result, the kinetic model was well fitted and additionally we confirmed that the estimated fitness values for 11 genotypes were considerably close to the experimentally measured ones (Ichihashi et al., 2015). In this sequel paper, with the theoretical basis of the deterministic kinetic model, we present the details of inference of the fitness values and putative appearance time points for the 91 genotypes. It may be possible to apply this methodology to other self-replicating molecules, viruses and bacteria.

1. Introduction

Many researchers have considered evolvable self-replicating RNA molecules as the origin of life (e.g. Eigen, 1992). Based on this concept, it is worth constructing an evolution system of the RNA molecules and observing the evolutionary dynamics in the system. We have established a translation-coupled RNA replication system within a cell-like compartment, where the RNA sequence used is a part of Qβ phage genome (Ichihashi et al., 2013). Through a serial transfer culture of the RNA molecules in the system, we observed Darwinian evolution of the RNA molecules. After random sampling at 11 time points and sequencing of them, we obtained a time series of occurrence frequencies of 91 individual genotypes. This revealed that a complex clonal interference occurs throughout the culture and the population is quite polymorphic one, which is called the "quasispecies" (Eigen, 1971, 1992). This phenomenon was reported in our previous paper (Ichihashi et al., 2015).

The aim of this paper is to present a method to estimate the fitness value and "putative appearance time point" of each mutant genotype by analyzing the time series of occurrence frequencies of them (Illingworth and Mustonen, 2012). The putative appearance time point is a certain time point at which a mutant genotype is supposed to appear in our deterministic kinetic model. Estimation of the fitness values and validation of the results were reported in our previous paper (Ichihashi et al., 2015). This sequel paper

* Corresponding author at: Department of Bioinformatic Engineering, Graduate School of Information Science and Technology Osaka University, Yamadaoka 1-5, Suita, Osaka, Japan.
E-mail address: tetsuyayomo@gmail.com (T. Yomo).

http://dx.doi.org/10.1016/j.jtbi.2016.04.014
0022-5193/

focuses on the details of the theoretical basis of the kinetic model and presents the procedure of estimating the fitness values and putative appearance time points. By estimating these parameters, we can depict snapshots of evolution process at any time point (Ichihashi et al., 2015). Furthermore, the analysis of the relationships between the estimated fitness values and genotype sequences is helpful to obtain a view of the fitness landscape of interest (Szendro et al., 2013) and to decompose the whole fitness into the site-fitness contributions of mutated individual residues (Aita et al., 2001). Our method may be applicable to more general cases (Long et al., 2015), particularly, cases of viruses or pathogens within a human host environment (Lee et al., 2009).

2. Methods

2.1. Model of population dynamics

Consider that evolvable self-replicating molecules with population size M are cultivated in a reactor, where M (the number of molecules) is fixed throughout the culture and $M=10^8$–10^{12} for real experiments. As an initial condition, the population contains a single master genotype that takes the largest mole fraction (=relative frequency). Mutant genotypes appear one after another through the evolution process. We focus on "major genotypes" whose mole fractions increase significantly larger than a certain detectable limit (we used 0.01 in mole fraction in this study) and neglect other minor genotypes. Let $N+1$ be the number of the major genotypes identified throughout the evolution process. Each genotype is represented by the serial number i ($i=0, 1, 2, \ldots, N$). For example, the single master genotype in the initial state is represented by $i=0$. Let k_i be the propagation rate constant of a genotype i. In this paper, we call the k_i the "fitness" of the genotype i for convenience. Let $x_i(t)$ be the mole fraction of a genotype i at time point t. Note that $\sum_{i=0}^{N} x_i(t)=1$ holds anytime. Let $D(t)$ be the dilution rate at time point t.

Hereafter, we assume that, from a viewpoint of the chemical reaction kinetics, the population dynamics is described as the following deterministic differential equations:

$$\frac{dx_i(t)}{dt} = (k_i - D(t))x_i(t) \quad (i=0, 1, 2, \ldots, N) \tag{1}$$

with $\sum_{i=0}^{N} x_i(t)=1$ and $D(t)=\sum_{i=0}^{N} k_i x_i(t)$. We define the "putative appearance time point" of a genotype i, denoted by τ_i, as a certain time point at which the genotype i is supposed to appear in the deterministic kinetic model. We assume that the occurrence frequency of each mutant genotype starts from a single molecule. The ith genotype participates in the competition from time point τ_i. Then, for all i's except $i=0$, $x_i(t)=0$ when $t<\tau_i$, and $x_i(t)=1/M$ when $t=\tau_i$. The value of τ_i is probabilistically given because mutational events occur randomly.[1] It should be noted that the τ_i does not necessarily mean the "true" appearance time point, because stochastic effects (genetic drift) due to small number of molecules make the phenomenon more complicated (Rouzine et al., 2001). This issue will be discussed later.

The solution of the above equations is given by

$$x_i(t) = \frac{A_i e^{(k_i-k_0)t} U\left(t-\tau_i\right)}{\sum_{j=0}^{N} A_j e^{(k_j-k_0)t} U\left(t-\tau_j\right)} \quad (i=0, 1, 2, \ldots, N), \tag{2}$$

where A_i is the following quantity as a function of the given set $\{k_i-k_0, \tau_i | i=1, 2, \ldots, N\}$:

$$A_i = \begin{cases} 1, & \text{for } i=0 \\ \dfrac{1}{M}\sum_{j=0}^{N} A_j e^{-(k_i-k_j)\tau_i} \widetilde{U}(\tau_i-\tau_j), & \text{for } (i\geq 1), \end{cases} \tag{3}$$

and $U(\,)$ and $\widetilde{U}(\,)$ are defined as the following step functions of z,

$$U(z) \equiv \begin{cases} 0, & (z<0), \\ 1, & (z\geq 0), \end{cases} \quad \widetilde{U}(z) \equiv \begin{cases} 0, & (z\leq 0), \\ 1, & (z>0). \end{cases} \tag{4}$$

In the above equations, $\tau_0 = -\infty$ for mathematical convenience. Eqs. (2) and (3) are derived in Appendix A. By using Eq. (3) recursively, the values of A_i's are determined in order of appearance in the evolution process. Eq. (2) with Eq. (3) states that the time series of the mole fraction $x_i(t)$ is governed by the parameter set $\{k_i-k_0, \tau_i | i=1, 2, \cdots, N\}$.

If an initial population contains only a single genotype $i=0$, it is obvious that $\tau_i>0$ holds for all i's except $i=0$. If an initial population contains other different minor genotypes (e.g. $i=1,2$), $\tau_i<0$ holds for these initial genotypes (e.g. $\tau_1, \tau_2<0$). In this case, the values of A_i's for these initial genotypes correspond to their mole fractions at the initial time point $t=0$ approximately (e.g. $x_1(0)\approx A_1$ and $x_2(0)\approx A_2$).

2.2. Inference of fitness values and putative appearance time points

For several observation time points, by random sampling of genotypes from the evolving population, the occurrence frequency of each genotype can be counted. Since the sampling size at each time point is considerably small in comparison with the whole population, only the major genotypes are identified. Let $N+1$ be the number of the identified genotypes (including the start genotype) throughout the experimental evolution. Consider that the values of $x_i(t)$ ($i=0\sim N$) are experimentally measured for several observation time points denoted by t_s ($s=1, 2, \ldots, S$).

We estimated the parameters k_i-k_0 and τ_i ($i=1, 2, \ldots, N$) in Eq. (2) by minimizing the following sum of the "Kullback–Leibler divergence" (Kullback and Leibler, 1951):

$$KL \equiv \sum_{s=1}^{S}\sum_{i=0}^{N} x_i^{\exp}(t_s) \log \frac{x_i^{\exp}(t_s)}{x_i(t_s)}, \tag{5}$$

where $x_i(t)$ represents the kinetic model given in Eq. (2) and $x_i^{\exp}(t_s)$ is the experimentally measured value. Namely, $\{k_i-k_0, \tau_i | i=1, 2, \ldots, N\} = \arg\min KL$. This fitting procedure is quite reasonable because the Kullback–Leibler divergence is typically used as a distance-like measure between two probability distributions (the mole fraction can be regarded as the probability distribution)[2].

The minimization of Eq. (5) was conducted by the steepest descend method starting from the following initial conditions: $\tau_i = T_i - 4 - 13.8/k_i$ and $k_i = 0.05\times i$, where T_i is defined as the time point at which each genotype i was firstly observed by random sampling and sequencing with sample size of 200–300 at each time point. The relationship between the initial values of k_i-k_0 and τ_i is shown in Fig.S2(a) in Supplemental materials.

The minimization was implemented by the C program running in PC cluster Express5800/120Rg-1 × 128 nodes (48GFLOPS, Intel Xeon 3 GHz (Woodcrest) 2CPU 4 core, 16 GB memory per node). The operating system is SUSE Linux Enterprise Server 10. The runtime was within 10 min.

[1] This is the large difference from the deterministic scheme of the original quasispecies theory (Eigen, 1971, 1992).

[2] Other studies used alternative measures such as the sum of the square of $(x_i(t_s)-x_i^{\exp}(t_s))/x_i^{\exp}(t_s)$ (e.g. Matsubara et al., 2006).